EXHIBIT 66

Anna's Archive Processing Pipeline

POC: Viktor Kerkez

What?

The project aims to scrape and process the entirety of Anna's Archive Containers (AAC), which standardizes releases from the world's largest shadow libraries.

Anna's Archive is an open-source and open-data initiative. They intend to systematically organize, enrich, and make the data more accessible for various uses, including research, educational purposes, and the preservation of human knowledge.

They achieve this by:

1. Mirroring existing open-data shadow libraries (like Sci-Hub and Library Genesis).
2. Helping out shadow libraries that want to be more open but don't have the time or resources to do so (like the Libgen comics collection).
3. Scraping libraries that do not wish to share in bulk (like Z-Library).

Pros:

1. Distributed through torrents, though with the possibility of other distribution methods (e.g. IPFS).
2. They do incremental releases / appendable releases.

Cons:

1. They don't care about files being easy to navigate manually on disk or searchable without preprocessing.
2. They don't care about being directly compatible with existing library software.
3. They don't expect the files to be usable without significant technical knowledge and commitment.

Contents

Dataset	Dataset Size	Dataset #Files	% Mirrored by AA	AA Size	AA #Files
libgen.rs	77.1 TB	6,940,937	93.49%	72.08 TB	6,488,734
Sci-Hub	87.2 TB	97,847,480	88.48%	77.15 TB	86,577,407
libgen.li	255.2 TB	17,417,854	83.04%	211.92 TB	14,463,960

Z-Library	102.8 TB	14,947,832	99.91%	102.71 TB	14,934,827
Internet Archive (CDL)	219.6 TB	7,919,904	66.79%	146,67 TB	5,289,624
Total (Deduplicated)	**590.5 TB**	**127,586,404**	**86.71%**	**512.02 TB**	**110.628.895**

Details about the content

Source	Metadata	Files
Libgen.rs	✅Daily HTTP database dumps.	✅Automated torrents for Non-Fiction and Fiction ❌❌ Anna's Archive manages a collection of book cover torrents.
Sci-Hub / Libgen "scimag"	❌Sci-Hub has frozen new files since 2021. ✅Metadata dumps available here and here, as well as as part of the Libgen.li database (which we use).	✅Data torrents available here, here, and here. ❌Some new files are being added to Libgen's "scimag", but not enough to warrant new torrents.
Libgen.li	✅Quarterly HTTP database dumps.	✅Non-Fiction torrents are shared with Libgen.rs (and mirrored here). ❌ Fiction collection has diverged but still has torrents, though not updated since 2022 (we do have direct downloads). ❌❌ Anna's Archive manages a collection of comic books and magazines. ❌No torrents for Russian fiction and standard documents

		collections.
Z-Library	✗No metadata available in bulk from Z-Library. □□ Anna's Archive manages a collection of Z-Library metadata.	✗No files available in bulk from Z-Library. □□ Anna's Archive manages a collection of Z-Library files.
Internet Archive Controlled Digital Lending	✓Some metadata available through Open Library database dumps, but those don't cover the entire Internet Archive collection. ✗No easily accessible metadata dumps available for their entire collection. □□ Anna's Archive manages a collection of Internet Archive metadata.	✗Files only available for borrowing on a limited basis, with various access restrictions. □□ Anna's Archive manages a collection of Internet Archive files.

Commented [1]: Just want to make clear this is 65% (per https://annas-archive.org/datasets/ia) of the Internet Archive controlled lending of books. It's not a replacement for getting Internet Archive web crawl, books without controlled lending, etc.

Commented [2]: Yes that's how I understood it from their docs.

Why?

Anna's Archive represents a pivotal resource in the digital age, encapsulating a vast array of knowledge spanning across multiple disciplines. The motivation behind this project is to enhance the accessibility and utility of this significant repository, ensuring that the wealth of information it contains is preserved and made available for future generations. Given its scale and the diversity of its contents, Anna's Archive is crucial for supporting a wide range of academic, scientific, and cultural research endeavours.
Incorporating this project into our datasets aligns with our top-line goal of advancing the democratization of knowledge and supporting open-source initiatives.

of New Tokens This Project Will Produce From

It is extreamly hard to estimate the number of tokens that can be extracted from this dataset because of it's extreamly high duplication of content. The same books can be found in different formats (pdf, epub, mobi, azw, etc.), in other editions, or just equivalent copies that, for formatting reasons, have different hash values.
As stated above:
Anna's Archive doesn't care about files being easy to navigate manually on disk or searchable without preprocessing.
Also, some data sources have a much higher duplication rate than others. The duplication rate of Z-lib is an order of magnitude higher than that of libgen.
The best proxy method to estimate the number of tokens is by comparing it to the libgen library we already have.

- Current libgen consists of libgen.rs + libgen.li + Sci-Hub, and at the moment of scraping, had **88,303,239** documents.
- Anna's Archive libgen contains **107,530,101** documents, which is **21.77%** more.
- In addition, Anna's Archive contains **20,224,451** more documents from Z-Lib and Internet Archive, which is **22.9%** of the current libgen.

So, in the best case, where there are absolutely no duplicate documents, we can expect to extract **44.67%** of the current libgen new tokens, which adds to **290.4B** new tokens.

Since these are considered extremely high-quality tokens, even with the high duplication rate, we can expect that getting Anna's Archive can substantially increase the same metrics the libgen dataset did.

Formatting

Ultimately, they settled on a relatively simple standard. It's pretty loose, non-normative, and a work in progress.

- **AAC.** AAC (Anna's Archive Container) is a single item consisting of **metadata** and optionally **binary data**, both immutable. It has a globally unique identifier called **AACID**.
- **Collection.** Each AAC belongs to a collection, which is a list of semantically consistent AACs. That means that if you make a significant change to the format of the metadata, then you have to create a new collection.
- **"records" and "files" collections.** By convention, it's often convenient to release "records" and "files" as different collections so they can be released at different schedules, e.g. based on scraping rates. A "record" is a metadata-only collection containing information like book titles, authors, ISBNs, etc, while "files" are the collections that contain the actual files themselves (pdf, epub).
- **AACID.** The format of AACID is this: aacid__{collection}__{ISO 8601 timestamp}__{collection-specific ID}__{shortuuid}.
    - {collection}: the collection name, which may contain ASCII letters, numbers, and underscores (but no double underscores).
    - {ISO 8601 timestamp}: a short version of the ISO 8601, always in UTC, e.g. 20220723T194746Z. This number has to increase monotonically for every release, though its semantics can differ per collection. They suggest using the time of scraping or generating the ID.

    - {collection-specific ID}: a collection-specific identifier, if applicable, e.g. the Z-Library ID. It may be omitted or truncated. Must be omitted or truncated if the AACID would otherwise exceed 150 characters.
    - {shortuuid}: a UUID compressed to ASCII, e.g. using base57. They currently use the shortuuid Python library.
- **AACID range.** Since AACIDs contain monotonically increasing timestamps, they can be used to denote ranges within a particular collection. They use this format: aacid__{collection}__{from_timestamp}--{to_timestamp}, where the timestamps are inclusive. This is consistent with ISO 8601 notation. Ranges are continuous and may overlap, but in case of overlap, they must contain identical records as the one previously released in that collection (since AACs are immutable). Missing records are not allowed.
- **Metadata file.** A metadata file contains the metadata of a range of AACs for one particular collection. These have the following properties:
    - Filename must be an AACID range, prefixed with annas_archive_meta__ and followed by .jsonl.zstd.
    - As indicated by the file extension, the file type is JSON Lines compressed with Zstandard.
    - Each JSON object must contain the following fields at the top level: **aacid**, **metadata**, **data_folder** (optional). No other fields are allowed.
    - Metadata is arbitrary metadata, per the semantics of the collection. It must be semantically consistent within the collection.
    - data_folder is optional and is the name of the binary data folder that contains the corresponding binary data. The filename of the corresponding binary data within that folder is the record's AACID.
    - The annas_archive_meta__ prefix may be adapted to the name of your institution, e.g. my_institute_meta__.
- **Binary data folder.** A folder with the binary data of a range of AACs, for one particular collection. These have the following properties:
    - Directory name must be an AACID range, prefixed with annas_archive_data__, and no suffix.
    - The directory must contain data files for all AACs within the specified range. Each data file must have its AACID as the filename (no extensions).
    - It's recommended to make these folders somewhat manageable in size, e.g. not larger than 100GB-1TB each, though this recommendation may change over time.

- **Torrents.** The metadata files and binary data folders may be bundled in torrents, with one torrent per metadata file or one torrent per binary data folder. The torrents must have the original file/directory name plus a .torrent suffix as their filename.

Processing

Processing of Anna's Archvie will follow the same procedure established by the pipeline that processed the libgen. With the difference of preprocessing documents to determine the duplicates before OCR-ing them to avoid wasted computing. Mainly because the libgen was processed using Nougat, which has high resource requirements and has to be run on GPUs.

Pipeline details can be found in the original libgen processing document. LibGen dataset [lab notebook]

Open Questions / Risks

- Open Questions:
  - How will incremental updates from Anna's Archive be managed and integrated?
  - What strategies can be employed to ensure long-term seeding and accessibility of torrents?
- Risks:

Redacted - Privilege

Document Comments
Total Comments: 2

Author: Kenneth Heafield
Date: 2/29/2024 5:15:00 PM
Range: Just want to make clear this is 65% (per https://annas-archive.org/datasets/ia) of the Internet Archive controlled lending of books. It's not a replacement for getting Internet Archive web crawl, books without controlled lending, etc.
Scope: Internet Archive Controlled Digital Lending·✓ Some metadata available through Open Library database dumps, but those don’t cover the entire Internet Archive collection.✗ No easily accessible metadata dumps available for their entire collection.🤷‍♂️💻 Anna’s Archive manages a collection of Internet Archive metadata.·✗ Files only available for borrowing on a limited basis, with various access restrictions.🤷‍♂️💻 Anna’s Archive manages a collection of Internet Archive files.··

Author: Viktor Kerkez
Date: 2/29/2024 5:22:00 PM
Range: Yes that's how I understood it from their docs.
Scope: Internet Archive Controlled Digital Lending·✓ Some metadata available through Open Library database dumps, but those don’t cover the entire Internet Archive collection.✗ No easily accessible metadata dumps available for their entire collection.🤷‍♂️💻 Anna’s Archive manages a collection of Internet Archive metadata.·✗ Files only available for borrowing on a limited basis, with various access restrictions.🤷‍♂️💻 Anna’s Archive manages a collection of Internet Archive files.··